An Efficient Time-Frequency Representation for Parametric-Based Audio Object Coding

Seungkwon Beack, Taejin Lee, Minje Kim, and Kyeongok Kang

Object-based audio coding can provide new music applications with interactivity. To efficiently compress a lot of target audio objects, a subband-based parametric coding scheme has been adopted for MPEG spatial audio object coding. In this letter, the time-frequency (T/F) subband analysis structure is investigated. A reconfigured T/F structure is also proposed to enhance the generating performance of sound scenes such as 'karaoke' and 'solo' play in interactive music scenarios. From the experimental results, it was confirmed that the proposed scheme remarkably improves the SNR and sound quality.

Keywords: SAOC, parametric audio coding, spatial cue.

I. Introduction

The meaning of an object in audio signal processing is the same as that of an audio track, but its usage can clearly be distinguished by regarding its application in such areas as interactive music. For interactive music, an audio track can be handled as an object by flexibly rendering the level and virtual position of the audio track to produce a specific sound scene according to the user preference. For instance, 'karaoke' and 'solo' instrument play are representative examples of possible preferred sound scenes in an interactive music service [1], [2].

The MPEG audio-subgroup recently developed a spatial audio object coding (SAOC) scheme for such an interactive music service [1]. SAOC is based on a parametric audio coding scheme with a high compression ratio, which is technically similar to parametric stereo [3] and MPEG surround [4]. Despite its high coding efficiency, however, such parametric coding scheme cannot fully support interactive functionality due to its intrinsic degradation produced by parametric modeling. The main problem of a parametric coding scheme is that each reconstructed audio track includes other tracks, which causes considerable interference. Thus, the degradation of sound quality is more easily perceived when the user wants to extremely render the level of a certain object, for example, a vocal track for enjoying a song in 'karaoke' mode.

In this letter, an alternative time to frequency (T/F) transform structure is proposed for analyzing spatial cues. The proposed structure also provides further fine resolution of spatial cues to relieve the level of interference signals from a desired track. It will be shown that finely-resolved spatial cues can improve sound quality, even in 'karaoke' and 'solo' modes, compared with SAOC.

II. Overview of MPEG SAOC T/F Structure

Figure 1 shows a schematic of the T/F structure of the current MPEG SAOC. A two-stage filter bank structure is adopted in SAOC. In the first stage, a 64-quadrature mirror filterbank (QMF) uniformly splits an input frame signal into 64 subband filtered signals, and then the filtering-based hybrid filterbank is subsequently applied into the 3 lowest subbands to make them split further into sub-subbands; the lowest is split into 6 sub-subbands and the other two bands are split into 2 sub-subbands. Consequently, 71 bands are finally obtained through the two-stage QMF structure. During the analysis stage of a spatial cue, the 71 bands are grouped into 20 (or 28) subbands with near alignment with the boundary of the equivalent rectangular bandwidth (ERB) [5], [6]. In other words, the maximum resolution of a subband for analyzing a spatial cue is restricted to the resolution of a 28-ERB-band frequency response at most.

Manuscript received Jan. 7, 2011; revised Mar. 22, 2011; accepted Mar. 30, 2011.

Seungkwon Beack (phone: +82 42 860 1745, email: skbeack@etri.re.kr), Taejin Lee (email: tjlee@etri.re.kr), Minje Kim (email: mkim@etri.re.kr), and Kyeongok Kang (email: kokang@etri.re.kr) are with the Broadcasting & Telecommunications Convergence Research Laboratory, ETRI, Daejeon, Rep. of Korea.

http://dx.doi.org/10.4218/etrij.11.0211.0007

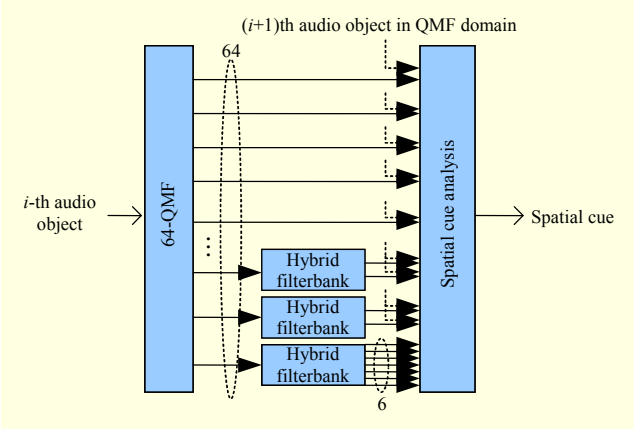

Fig. 1. Current MPEG SAOC T/F analysis structure.

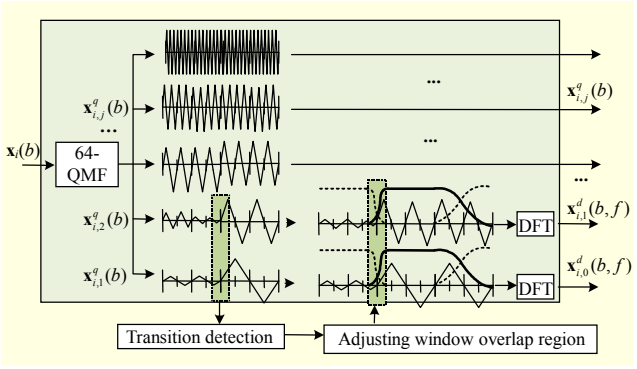

Fig. 2. Proposed T/F structure with temporal processing.

Even though these analysis band resolutions are limited by the QMF, it has been confirmed that the amount of 28-ERB-based subband processing is sufficient to successfully compress audio track signals with acceptable perceptual distortion. However, this is only perceptually guaranteed under the scenario in which all of the decoded track signals are simultaneously playing. For scenarios such as 'solo' and 'karaoke' modes, ERB-based subband processing does not guarantee that the perceptually acceptable quality will continue to be ignored. This situation is confirmed through a subjective listening test detailed in the evaluation section of this paper. To overcome these weaknesses, SAOC supports enhanced 'karaoke' and 'solo' modes by adopting residual coding tools, but this is outside of the technical scope of this letter.

III. Proposed Reconfigured T/F Structure

Before describing the proposed scheme, the common strength and drawback of the discrete Fourier transform (DFT) and QMF as subband analysis methods can be summarized as follows. The strength of DFT is that it can provide flexible frequency resolution for subbands within the sampling rate and transform size by flexibly grouping the frequency bins; however, its main drawback is that the temporal information cannot be estimated in the transformed domain, and thus temporal processing within the subband, such as decorrelation and the detection of a transient interval, requires an additional filtering operation and results in the use of a considerable amount of computational power. Conversely, the strength of QMF is that subband-based temporal processing can be easily achieved from the transformed domain because signals transformed by QMF are basically time-sampled versions through the filtering operation using a prototype low-pass QMF filter; however, QMF does not easily adjust the resolution of the subband because the resolution is determined by the prototype filter.

Our approach is basically designed to combine the strengths of both DFT and QMF and thus lessening the drawbacks of each. Figure 2 shows a schematic of the proposed T/F structure combined with the temporal processing of the transition coding by adjusting the overlap size of the analysis windowing. The procedure used in the proposed scheme is summarized below.

First, we apply a normal 64-QMF, and the lowest filtered band signals are then transformed to the frequency domain by DFT. The number of lower bands can be selected depending on the characteristics of the input signals. Also, the subband can be flexibly grouped in the DFT frequency domain. To apply the temporal coding scheme, the transition interval is searched in the QMF domain, and we adjust the analysis window shape (that is, the adjusting overlap region) to prevent a spread of quantization noise of the spatial cue into the adjacent frames during the synthesis process in the DFT domain.

A more detailed description of the proposed scheme can be described using a vector and matrix operation including a spatial cue analysis stage. The input signal frame, denoted as a vector, $\mathbf{x}_i(b) = [x_{i,j}(b), x_{i,j}(b-1), \ldots, x_{i,j}(b-(N-1))]^T$, is first applied into the forward QMF:

$$\mathbf{X}^q_i(b) = \mathrm{QMF}\{\mathbf{x}_i(b)\} = [\mathbf{x}^q_{i,1}(b), \ldots, \mathbf{x}^q_{i,M}(b)]^T, \qquad (1)$$

where i is the index of the audio track $(0 \le i \le M-1)$, j is the index of the QMF subband $(0 \le j \le 63)$, N is the frame size, and T is the transpose operator. The block index with increased step of block size is b. In (1), for example, b is increased by the step of N. The output matrix of QMF is $\mathbf{X}^q_i(b)$, which consists of vector $\mathbf{X}^q_{i,j}(b)=[x^q_{i,j}(b), \ldots, x^q_{i,j}(b-(L-1))]^T$ with length L calculated by N/M. As a temporal processing, the transient interval can be extracted from $\mathbf{x}^q_{i,j}(b)$ by adopting the common transient detection algorithm [3], [4].

After the forward QMF, the lower j' $(0 \le j' \le J < 64)$ $\mathbf{x}^q_{i,j'}(b)$ is combined with the previous frame vector in order to apply DFT, as denoted by

$$\mathbf{x}^{d}_{i,j'}(b,f) = DFT\left\{W \times \left[\mathbf{x}^{q}_{i,j'}(b-1)\mathbf{x}^{q}_{i,j'}(b)\right]^{T}\right\}, \quad (2)$$

where J is the maximum number of QMF-bands to apply (2). In (2), W is a common sine window matrix and is the operator of element-wise multiplication. The element of $\mathbf{x}^{d}_{i,j'}(b,f)$ is a frequency bin with a complex value, and f is a frequency bin index.

To estimate the object level difference (OLD), the power estimation in a DFT vector can be written as

$$Pow^{d}_{i,j',k} = \sum_{f=A_k}^{A_{k+1}-1} \left|\mathbf{x}^{d}_{i,j'}(b,f)\right|^{2} \quad (0 \le k \le K), \quad (3)$$

where K is the maximum number of DFT subbands to apply (3). In (3), A_k is the k-th parameter boundary within the j'-th QMF band. This means that one DFT vector from one QMF band can further split the subvector again into a sub-subband in order to estimate finer spatial cues, and the number of sub-subbands within $\mathbf{x}^{d}_{i,j'}(b,f)$ can be easily changed by regrouping its elements. In the remaining higher bands ($J < j$) of QMF, the power estimation is defined as

$$Pow^{q}_{i,p} = \sum_{j=B_p}^{B_{p+1}-1} \left|\sum_{n=0}^{n=L-1} \mathbf{x}^{q}_{i,j}(b-n)\right|^{2} \quad (0 \le p \le P), \quad (4)$$

where P is the number of the QMF-bands to apply (4). In (4), the QMF bands are grouped according to the boundary B_p among the QMF bands, which normally follows the boundary of ERB. From (3) and (4), the two OLD parameter types are estimated by

$$\Delta OLD^{d}_{i,j',k} = 10\log_{10}\left(\frac{pow^{d}_{i,j',k}}{pow^{d}_{i+1,j',k}+\alpha}\right), \quad (5)$$

$$\Delta OLD^{q}_{i,p} = 10\log_{10}\left(\frac{pow^{q}_{i,p}}{pow^{q}_{i+1,p}+\alpha}\right), \quad (6)$$

where $\Delta OLD^{d}_{i,j',k}$ is the fine spatial cue for the synthesis of the lower QMF bands, and $\Delta OLD^{q}_{i,p}$ is a normal cue for the synthesis of higher QMF bands. Therefore, it can be estimated that the total number of OLD sets is $J \times K + P$.

In the structure, it is possible that the numbers of K and J are flexibly adjusted according to the rendering scenario. For instance, if the input objects consist of a vocal track and other instrumental object tracks, and the decoder should support the playback scenario of either vocal 'solo' or 'karaoke' mode, the number of K for a certain j' could be increased to analyze the spatial cues with more sub-subbands. An increase in J results in an extension of the number of sets of K. These adjustments are more appropriate for enhancing the sound fidelity by removing the interference signals from the desired target vocal object.

Table 1. Legend and bitrates corresponding to evaluation system.

Type	Legend	Average bitrate/track
A	28-subband analysis based on SAOC	2.4 kbps
B	71-subband analysis based on SAOC	6.02 kbps
C	48-subband analysis based on proposal	4.21 kbps
D	68-subband analysis based on proposal	5.91 kbps

IV. Experimental Results

Objective and subjective measurements were conducted to confirm the effectiveness of the proposed method. In the objective measurement, five test items were used, where each item consisted of two audio objects, vocal and music recoding (MR) tracks. All of the items were sampled at 48 kHz with 16-bit quantization. The average segment SNR was adopted as an objective measurement, where a segment SNR is denoted as [7]

$$\text{segSNR}=20\log_{10}\left(\frac{\sum_{n=0}^{N-1} m(n)}{\sum_{n=0}^{N-1}(m(n)-\hat{m}(n))^{2}}\right), \quad (6)$$

where N is the length of a segment defined as 1,024, and $m(n)$ is an original mixed signal with original vocal $s_{\text{vocal}}(n)$ and MR $s_{\text{mr}}(n)$ from (7). Also, $\hat{m}(n)$ is a decoded mixed signal by (7), with a decoded vocal $\hat{s}_{\text{vocal}}(n)$ and MR $\hat{s}_{\text{mr}}(n)$:

$$m(n) = [1-\alpha \quad \alpha]\begin{bmatrix} s_{\text{vocal}}(n) \\ s_{\text{mr}}(n) \end{bmatrix}, \quad (7)$$

where $m(n)$ and $\hat{m}(n)$ are generated eleven times according to the increment of mixing parameter α with 0.1 steps ($\alpha = 0, 0.1, ..., 1$). When α=0, 'solo' mode can be evaluated, that is, $m(n)$ and $\hat{m}(n)$ are $s_{\text{vocal}}(n)$ and $\hat{s}_{\text{vocal}}(n)$, respectively. On the other hand, 'karaoke' mode can be generated using α=1. Four types of systems with different analysis structures for the spatial cues were evaluated, and bitrates are described in Table 1. From the bitrate comparison, the proposed method accommodates slightly increasing bitrate, but the following experimental results clearly show improved performance, whereas type B does not show any significant difference of performance. For type D, we intentionally extended the number of SAOC parameter bands to 71, which is the maximum number of allowed subbands in the current SAOC T/F structure. Table 2 shows the configuration of the subband layout used in system types A and B. The number of J for type A is 8, which indicates that DFT sub-subbands cover the 8×375-Hz bandwidth at a 48-kHz sampling rate, and the different number of sub-subbands is assigned at each j'.

Table 2. Configuration of subband layout of types C and D.

Type	J	K	P
C	5	16 (j'=0), 8 (j'=1, 2), 4 (j'=3, 4)	8
D	8	16 (j'=0, 1), 8 (j'=2, 3), 4 (j'=4, 5), 2 (j'=6, 7)	8

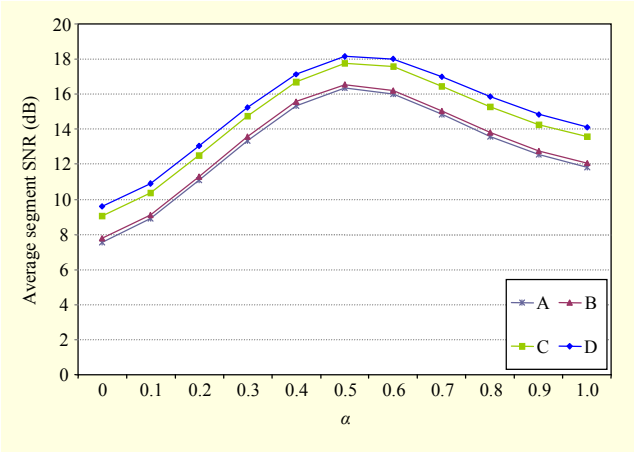

Fig. 3. Average segment SNR performance according to α.

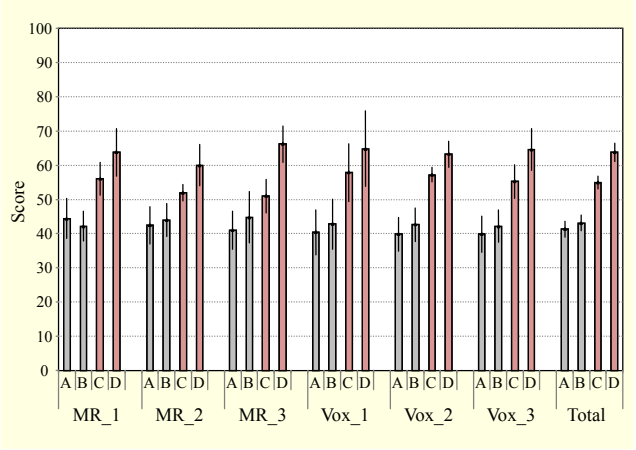

Fig. 4. Subjective score with 95% confidence interval.

Figure 3 shows the average segment SNR from each of the five items. The x-axis denotes the α value, and the y-axis denotes the dB of the average segment SNR. From the figure, it can be known that the SNR values are relatively lower in the mixed 'solo' and 'karaoke' cases. Also, type A nearly overlaps its performance with type B. This indicates that the performance does not need to be improved, despite an increase in the number of spatial cues within the current MPEG SAOC T/F structure. Types C and D based on the proposed scheme consistently show a superior performance of about 2 dB, which confirmed that the proposed fine parametric bands effectively works on the objective performance

With confidence that an objective performance gain can affect the sound quality, a subjective listening test was conducted. In the evaluation, MUSHRA was selected as the listening test methodology [8]. In clearly showing the comparison of the evaluation codec, the hidden reference and anchor need not be included in Fig. 4. Regarding 'solo' and 'karaoke' play, three decoded vocal tracks and three decoded MR tracks were evaluated as the test items, and six audio experts participated in the test. Figure 4 shows the test results with a 95% confidence interval. The mean score of systems C and D consistently demonstrate a higher value than those of A and B. Some items show a significantly better performance within the confidence interval. The total score indicates that proposed systems are significantly better than the current SAOC-based system.

V. Conclusion

An alternative T/F structure was proposed for subband-based parametric coding. The proposed structure is more appropriate for handling audio signals as objects with minor increase of bitrates. The main advantage of the proposed structure comes from combining the strengths of DFT and QMF. Regarding the complexity issues, it can be expected that the FFT will reduce computational power of the proposed scheme instead of the filtering operation of hybrid filterbanks of SAOC. The evaluation results confirmed the improved performance in both objective and subjective measurements. The proposed coding scheme is expected to be helpful in accelerating new music applications with interactivity.

References

[1] ISO/IEC 23003-2:2010, "Part 2: Spatial Audio Object Coding," International Standard, Oct. 2010.

[2] T. Lee et al., "A Personalized Preset-based Audio System for Interactive Service," AES Convention, Oct. 2006.

[3] ISO/IEC 14496-3:2001, "Parametric Coding for High Quality Audio," Dec. 2003.

[4] ISO/IEC 23003-1:2007, "Part 1: MPEG Surround," International Standard, Jan. 2007.

[5] C. Faller and R. Baumgarte, "Binaural Cue Coding-Part II: Schemes and Application," *IEEE Trans. Speech Audio Proc.*, vol. 11, no. 6, Nov. 2003.

[6] S. Beack et al., "Angle-Based Virtual Source Location Representation for Spatial Audio Coding," *ETRI J.*, vol. 28, no. 2, Apr. 2006, pp. 219-222.

[7] 3GPP TS 26.290, Extended Adaptive Multi-Rate-Wideband Codec (AMR-WB+): Transcoding Functions.

[8] ITU-R Recommendation, *Method for the Subjective Assessment of Intermediate Sound Quality* (*MUSHRA*), ITU, BS. 1543-1, Geneva, 2001.